Plan for Visual Species Similarity:

Goal:

The goal for this task is to find a robust and quantitative manner to create a list of the most visually similar species to a given species(for us Antelopes).

Background:

The three concepts important to understanding this plan are: self-supervised learning/contrastive learning, determining the “limb” ratios of an animal from its key points, and determining the variance of labeled key points around some centroid of an animal. Below are resources to understand what it is and why it is important.

Self-supervised Learning/Contrastive Learning:

What:

Contrastive Learning

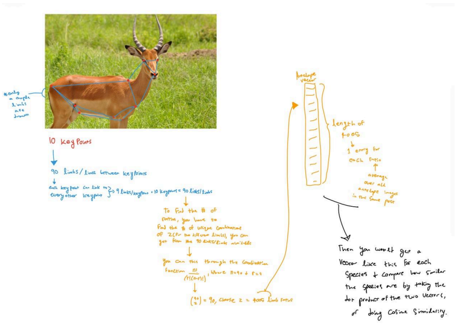
Self-Supervised Learning

Why:

These concepts are important because they allow AI models to generate very descriptive embeddings of animals and based off of core underlying features in an animal. It seems they generally allow the AI model to more deeply understand what is in the image and encode that information compactly into an embedding.(Add more)

Key Point “Limb” Ratios:

What:


*only a couple limbs are drawn
10 Key Points
90 limbs / lines between keypoints
each key point can link to every other keypoint → 9 links/keypoint • 10 keypoints = 90 links/limbs
To find the # of ratios, you have to find the # of unique combinations of 2(for the different limbs), you can get from the 90 links/limbs available
You can this through the Combination function $\frac{n!}{(r!(n-r)!)}$, Where n=90 + r=2
$\binom{90}{2}$ = 90, choose 2 = 4005 limb ratios
Antelope vector
length of 4005
1 entry for each ratio
average over all antelope images in the same pose
Then you would get a vector like this for each species + compare how similar the species are by taking the dot product of the two vectors, or doing cosine similarity.

Why:

Key point “limb” ratios could be an important metric for separating species based on key points as while, the key points relative to the animal in a species will change from image to image, the overall ratios between the “limbs”/lines connecting each key point should all hover around some mean, and have some(hopefully low) variance for a single species.

Key Point Variance Around Centroid:

What:

Still Need more explicit definition of centroid. Could be centroid around each key point or Centroid of image or body of animal.

Why:

Measuring the variance around a centroid of the animal and its key points would allow us to determine the range of motion of the animals in each species as even if two species are similar looking and have similar key point “limb” ratios, if they have wildly different ranges of motion, than we should not consider them very visually similar(assuming our downstream task is keypoints).

Outline:

There are currently three proposed approaches:

- The first uses two handcrafted features(key point “limb” ratios and key point variance around the centroid) to create a vector from those features for each image, and then performs clustering based on the cosine similarity of those vectors to determine a list of the most visually similar species to Antelopes.
- The second approach just uses a pre-trained self-supervised/contrastive-learning model like SimplCLR or Dino-v2 to get good-quality embeddings of the images and then uses these embeddings to perform clustering and so determine a list of the most visually similar species to Antelopes
- The third approach attempts to combine the first two approaches by using the two hand-crafted features from the first approach as constraints in the loss function used to fine-tune a model like SimplCLR or Dino-v2, as including these features which indicate similar key points could help the model prioritize not only similar looking species but similar looking species with the same type of features that affect key point accuracy(range of motion and “limb” ratios)

Thus, we start by experimentally setting up and conducting approaches 1 and 2, and then if we have time we can combine them to execute approach 3.

Approach #3(need to modify and make more concise):

I propose that we can determine the visual similarity between animal species by a multi-phase process:

1. Select a pre-trained self-supervised/contrastive learning model that can generate good-quality embeddings(small vectors that encode all the information in a given image)
2. Then to get the list of visually similar species, we can use a clustering algorithm to cluster all of the embeddings from the AP10K images(once we pass them through the model) + some of our real camera-trap images(not sure if we should actually do this, but if so could just add them to categories with AP-10K)
3. Then we can perform two different experiments and then compare to just the base self-supervised/contrastive learning model by determining which list of similar species(If used to train an RTMPose model on that list from the AP-10K dataset, and then tested on antelopes from AP-10K + some of our images)
    a. Fine-tune the model on some of our antelope images: I am not sure if this will do much, but if we run the same training process but just add some of our real-world antelope images(or other of our camera trap images, like of primates, or buffaloes) it could make the embeddings better?
    b. Fine-tune the model on the AP-10K dataset but add a constraint to the loss function the model is trying to optimize for:
        i. The model should try to make the difference between its embedding to another image closer or farther based on the measure of the cosine similarity between two vectors(or one if we concatenate both) of “limb” ratios and variance around a centroid. We should also add a hyper-parameter to this term so it acts more like regularization, so we can

manually control how much this condition affects the loss we are trying to optimize for.

Additionally, the hypothesis for why adding these extra conditions to the loss function could be beneficial is that we know the downstream task of identifying visually similar species, so it could be beneficial to curate the list of species, to prioritize ones that have “similar” keypoint definitions. For now, we have determined two metrics of similar key points are “limb” ratios and variance of the key points in an animal around a centroid/central point(trying to approximate the range of motion of the animal).

We could also add a possible ablation study where we only use one of the two conditions listed above to fine-tune the model, but not both, and then compare the results.